计算模型视角下信任形成的心理和神经机制

—— 基于信任博弈中投资者的角度*

高青林[1,2] 周媛[1,2]

([1]中国科学院心理研究所行为科学重点实验室，北京 100101)

([2]中国科学院大学心理学系，北京 100049)

摘 要 人际信任渗透在社会交互的各个方面，是促进和维持合作的重要基石。以往研究者借助信任博弈范式，主要探讨了人际信任的理论模型、生物基础和影响因素等方面。近年来，研究者开始将计算模型应用于信任博弈的数据分析中，深入挖掘人际信任行为背后的心理机制，将计算模型与神经影像技术结合，加深对信任行为背后脑机制的理解。目前将计算模型应用于信任博弈范式中的研究主要针对“信任是如何形成的”这一科学问题，未来要进一步发展计算模型方法，结合非侵入性脑刺激技术，应用于精神疾病人群中，以深入理解正常和异常信任形成的心理和神经机制。

关键词 人际信任; 信任博弈; 计算模型; 功能磁共振成像

1 引言

信任（trust）是经济以及社会生活中的一种润滑剂(Snijders & Keren, 2001)，也是维系社会关系的粘合剂(Wilson & Eckel, 2006)。人际信任 (interpersonal trust) 作为最复杂的社会技能之一，在社会交互中起着重要的作用(Fett et al., 2014) 。尽管人际信任的概念存在不同的表述，但其核心是指人们基于对他人行为的积极预期（比如在可能合作也可能竞争的情况下，预期对方会与自己合作），愿意将自己处于风险境遇中的一种心理状态 (Krueger et al., 2007; Rotter, 1967)。从定义可以看出人际信任的关键特征是：对他人意图持有积极预期；使自己陷入风险或劣势。经济学家在经济博弈理论的框架下，将人际信任从复杂的定义背景中抽离出来，并保留了信任的重要特征，进而操作成信任博弈范式(Trust Game, TG)。该范式被广泛地应用到人际信任的研究中。既往研究者从神经递质、激素等分子层面，决策推理等认知层面，以及脑区、脑网络等脑功能层面探究了人际信任的生物基础，并提出了各种理论模型来理解人际信任背后的心理和神经机制(Krueger & Meyer-Lindenberg, 2019;

1投稿日期：2020 年 4 月 18 号

*中国科学院心理研究所项目(E0CX163008)资助。

通讯作者：周媛，E-mail: zhouyuan@psych.ac.cn

Riedl & Javor, 2012; Tzieropoulos, 2013; 陈欣, 叶浩生, 2009; 陈瀛 等, 2020; 史燕伟 等, 2015; 张宁 等, 2011; 张蔚 等, 2016)。这些研究对于理解“人们为什么以及何时选择信任或不信任”这一科学问题提供了答案(张蔚 等, 2016)，但是传统的研究方法并不能回答“信任是如何形成的”这一科学问题。重复信任博弈范式的产生以及计算模型研究方法的应用，使得回答这一问题成为可能。

近年来，计算模型(computational modelling)方法因其严谨科学的量化方式，以及可以揭示行为和脑活动背后隐藏的动态心理过程的优势，而被越来越多地应用在决策领域中。这为加深理解行为背后的心理机制和神经基础提供了一种新的思路(Montague et al.,, 2012; Read Montague, 2018)。同时，这种基于数据的科学量化的方式不仅可以检验模型本身的好坏，也可以通过模型比较得出哪种模型能更好地解释和预测心理现象(Cheong et al., 2017)。其中，博弈范式是计算模型研究中常用的研究范式。在重复博弈中，人们需要推断对手当时的心理状态，而且这种推断具有递归性(recursive)，即一种重复循环的因果关系。这种递归性正是许多计算模型建立的核心思想，因而越来越多的研究将计算模型应用在包括信任博弈在内的多种博弈范式中，以此来探讨各种心理现象背后的内在机制(Ray et al., 2009)。

本文首先介绍了信任博弈范式，然后介绍了计算模型的概念，围绕“信任是如何形成的”这一科学问题介绍了计算模型在人际信任行为学研究和脑功能影像学研究中的应用，从而归纳了信任形成的心理和神经机制研究进展，最后针对目前研究的不足，为进一步探讨信任形成机制提供了新的思路。

2 单次和重复信任博弈

信任博弈是研究人际信任的最常用范式。在经典信任博弈范式中(Kreps, 1990)，由两个实验参与者分别扮演投资者(investor)和被信任者(trustee)的角色。博弈初始，双方拥有相同数额的金钱。首先由投资者在信任（把钱全部交给对方）和不信任（把钱全部保留）之间进行决策；如果选择信任，那么投资者投资的钱会翻倍（通常是三倍）给到被信任者；如果选择不投资，则本次游戏结束，双方钱数不变。然后由被信任者在互惠（把一半的钱返还给对方）和不互惠（把钱全部保留）之间进行决策。如果选择互惠，则双方最后的获益均是在原始金额上翻倍；如果选择不互惠，则被信任者将获得三倍于原始金额的钱数，而投资者的获益为零。通过该实验范式可以将信任(trust)量化为投资者的决策，将能否值得被他人信赖(trustworthiness)量化为被信任者的决策。Berg 等(1995)在此基础上修改形成了标准信任博弈范式。与经典信任博弈不同之处在于，在标准信任博弈范式中，投资者和被信任者可以自愿决定给出或者返还给对方多少钱，而不是全部给出或者全部保留。由此可以测量不同水平的信任和互惠。采用这种范式，Berg 等(1995)发现即使与陌生人只进行一次交易，人们还是会选择信任和互惠。而且这一研究结果也被后续多项研究证实(Declerck et al., 2013; Johnson & Mislin, 2011)。也有研究者根据特定的实验目的对该范式进行了变式，

例如在被试决策前允许博弈双方进行一分钟的语言交流，投资者在决策前会收到被信任者承诺返还的纸条，与真实的或者计算机模拟的被信任者博弈，以及与社会地位水平不同的被信任者博弈等等。以此来探究人们在信任博弈中做出信任行为的影响因素(Ben-Ner et al., 2011; Blue et al., 2020; Ma et al.,2015; Tzieropoulos, 2013)。

在经典的信任博弈中，同一对玩家之间的博弈是单次的(single-round)，但是现实生活中的社会交互很少只进行一次。因此，研究者进一步提出了重复信任博弈范式(Repeated Trust Game, RTG)，即同一对玩家之间连续进行多次信任博弈，玩家在博弈中可以及时获得反馈以此来调整下一次的决策（图 1）。不同于单次信任博弈，重复信任博弈中决策双方都承担对方可能不会把钱给自己的风险。因此，不仅投资者做出的信任行为会受到被信任者返还金额的影响，被信任者也需要考虑投资者的行为（在单次信任博弈中，被信任者并不会考虑投资者的行为）。可以看出，在重复信任博弈中玩家的行为不同于单次博弈。有研究发现，为了让投资者投资更多的金额，在重复信任博弈中被信任者返还的金额比在单次博弈中多(Cochard et al., 2004)；而且参与者做出的信任和互惠的决策呈单调递减趋势，并出现结尾效应(endgame effect)，即在博弈接近结尾的阶段，参与者选择信任以及互惠的决策骤然下降(Anderhub et al., 2002; Keser, 2003)。与单次信任博弈相比，重复信任博弈中涉及到学习、推理以及策略更新等多个认知过程，这为研究信任形成的过程提供了一种更生态的实验范式，也使得在社会交互情境下引入强化学习等计算模型成为了可能(Anderhub et al., 2002; King-Casas et al., 2005)。

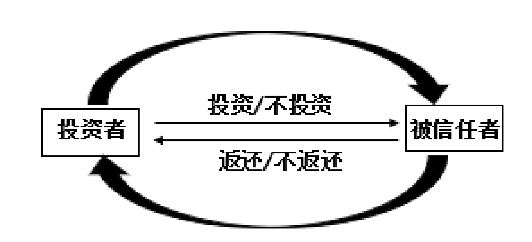


图 1. 重复信任博弈范式的示意图

3 计算模型在信任博弈中的应用

3.1 计算模型概述

计算模型是用抽象的数学表达式来刻画人类社会互动中学习以及决策的动态变化过程(Hackel & Amodio, 2018)，可以在行为或者脑活动的基础上刻画背后隐藏的动态变化的心理过程(Montague, 2018)。计算模型不仅可以基于行为探讨心理现象的动态变化过程，还可以与脑影像技术结合起来探讨心理现象背后的脑机制。其中，在当前发展较快的是计算模型与脑影像技术相结合的方法，例如：基于模型(model-based)的功能磁共振成像技术(functional Magnetic Resonance Imaging, fMRI)。fMRI 技术通过衡量 BOLD(Blood Oxygen Level Dependent)信号的变化来测量被实验刺激所诱发的脑活动，即某个脑区的 BOLD 信号增强代表该脑区被激活。为研究脑和行为的关系，在传统的 fMRI 研究中，研究者通常将

BOLD 信号与被试的准确率、反应时等行为指标建立起联系，从而得出某种行为倾向与脑功能活动之间的关联(Engelmann, 2010)。而基于模型的 fMRI(model-based fMRI)研究，可以通过模型计算将一些不能从实验范式中直接观察到的内部变量（例如奖赏预期偏差、学习速率）从行为数据中提取出来，模拟产生某种行为现象背后的复杂认知过程；再将这些变量或者模型参数与实验刺激诱发的 BOLD 信号建立起联系。由此可建立行为、认知以及脑功能活动之间的联系，从而更好地理解行为背后的脑机制 (Charpentier & O’Doherty, 2018; O'Doherty et al., 2007)。

当前用于人际信任研究的计算模型可以分为两类，即基于结果的模型(outcome-based model)和基于意图的模型(intention-based model) (McCabe et al.,2003)。基于结果的模型认为在信任博弈中，人们对于意图的推断不重要，重要的是个体从互动中获得的反馈结果，即在交易中人们主要看重的是自己的收益；而基于意图的模型则强调人们推断对方的意图在博弈决策的过程中更重要，即在互动中人们依据对方的意图作相应的决策。目前用在信任博弈中基于反馈结果的模型主要是强化学习模型(Cisler et al., 2015; Fouragnan, 2013; Radell et al., 2016)，基于意图的模型主要是贝叶斯模型(Jung et al., 2017; Moutoussis et al., 2014; Ray et al., 2009)。已有的使用强化学习模型探究人际信任的研究主要解决了先验可信度如何促进信任形成的问题，而贝叶斯模型在人际信任研究中主要解决的是对方意图的推测如何促进信任形成的问题。由于信任博弈中只有投资者行为反映人们的信任决策，而且目前使用计算模型探究信任博弈中信任行为相关问题时，均只对投资者行为进行分析，所以本文从投资者的角度分析信任博弈中信任行为计算模型的研究结果。

3.1.1 强化学习模型

强化学习模型(Reinforcement learning，RL)是最常用于经济决策中心理和神经机制建模的计算模型，用来解决人们如何从与环境多次互动产生的反馈中进行学习的过程(Read Montague, 2018)。该模型假设个体与外界环境互动的过程是马尔可夫决策过程(Markov decision process, MDP)。在该过程中，包括环境状态(State, S)，个体行动(Action, A)以及将二者联系起来的状态转移概率(Transition Probabilities, P)和奖赏(Reward, R)。其中状态指个体当时所处的位置。状态决定了个体能采取的行动有哪些，状态转移概率表明采取某种行动后，从一种状态转变到另一种状态的可能性(Puterman, 1995)。如图 2 所示，在 t 时刻主体(agent) 感知当前所处的状态 S_t 和当前所获奖赏 R_t，之后采取行动 A_t。主体采取的行动会引发 t+1 时刻环境所处的状态 S_{t+1} 以及奖赏值 R_{t+1}，而在该行动下环境从一种状态转变成另一种状态的可能性即为状态转移概率 $P(S_{t+1} |S_t , A)$ (Fouragnan, 2013）。强化学习模型认为个体通过习得当时所处环境状态下的行为与反馈结果之间的关系，在预期偏差 (Prediction Error)的基础上更新不同状态下某种行为的期望效用值，在最大化自己奖赏值的原则下做出适应性行为。该模型中，预期偏差指的是预期值和实际观测值之间的差距；学习速率反映

了个体对于结果反馈赋予的权重，用来衡量个体更新期望效用值的速度，值越大表明个体对反馈结果赋予的权重越大，更新期望效用值的速度越快(Claus & Boutilier, 1998)。

强化学习依据是否存在先验模型分成无模型(model-free)和基于模型 (model-based)两类(Montague et al., 2012)。无模型的强化学习理论认为，个体依据“试错”(trial-and-error)原则进行决策，即个体仅会依据过去习得的结果进行决策，类似于刺激-反应(stimuli-response)的习惯化(habitual)行为。最常用的模型是 Rescorla–Wagner（RW）模型。而基于模型的强化学习理论认为，个体会基于反馈形成一个自身对外部环境理解的内部模型，该模型是个体对外部世界的内部表征，个体在此基础上完成目标导向性行为(goal-directed)(Daw & Doya, 2006)。二者的主要区别在于是否有内部模型，基于模型的强化学习因其具有内部模型，所以加工反馈结果的方式更灵活，在该模型下的个体适应环境变化的速度更快。

对于信任博弈而言，投资者获得的反馈信号来自于被信任者是否返还金钱或者返还金钱数的多少。在无模型的强化学习假设中，不管外界有没有给出被信任者是否值得相信的线索，投资者都只会根据观察得到的对方可信度水平进行决策；而在基于模型的强化学习中，则假设投资者会先根据对方名誉线索对被信任者形成一个是否可信的先验期望，然后基于这种先验期望更新后续的预期偏差(Fouragnan, 2013)。

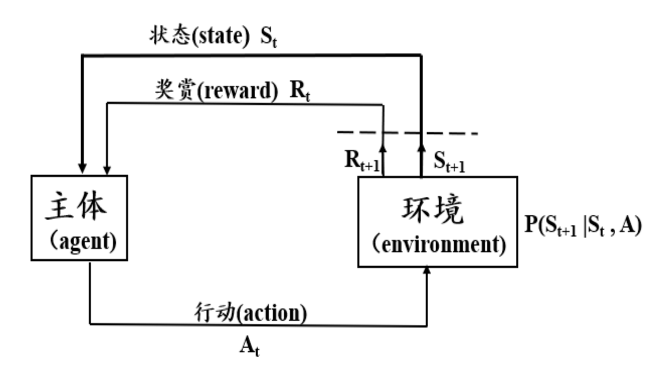


图 2. 强化学习模型框架图
资料来源：Fouragnan. (2013)

3.1.2 贝叶斯模型

强化学习模型是基于经典经济学家提出的个体是完全理性的假定，但违背这一完全理性假定的发现不断被报道(Fehr & Schmidt, 2005)。并且，强化学习模型假定个体需要通过马尔可夫决策过程获取环境中所有可能的状态，但是在实际社会互动中，个体所处的环境是不确定的、是部分可观测的。因此，研究者提出了基于部分可观测马尔可夫决策过程(partially observable MDP, POMDP)的贝叶斯模型(Bayesian model)。该模型认为个体是有限理性的。在社会互动前，个体对外界环境所处状态会有某种偏好，这种偏好即个体的先验信念(prior belief)。在互动过程中，个体会基于先验信念和环境反馈更新自己的先验信念，这种更新后的信念即个体的后验信念(posterior belief)。个体会基于后验信念做出适应性的决策行为。该类模型一般采用概率分布来表示信念(Kaelbling et al., 1995)。如图 3 所示，用先验概率分布 Pr 表示主体(agent)在加工外在信息前的先验信念，用后验概率分布 P 表示主

体在加工信息后对于环境所处状态形成的后验信念。其中，t 时刻后验信念的形成是基于该时刻下的先验信念 Pr，t 时刻互动行为观测集合 O_t 和奖赏集合 R_t 共同作用而成。主体在后验信念的基础上采取行动 A_t。主体采取的行动同样会引发 t+1 时刻的互动行为观测集 O_{t+1} 以及所获的奖赏集合 R_{t+1}，基于 O_{t+1} 和 R_{t+1}，主体的先验以及后验信念会得到进一步的更新，进而根据新的后验信念，在 t+1 时刻做出新的行动。贝叶斯推断模型与强化学习模型区别在于，后者是值函数随时间进行迭代，而前者是信念分布（先验信念与后验信念）随时间进行迭代(Friston et al.,2013)。

当马尔可夫决策过程中的状态是信念状态，同时这些信念状态是不确定且部分可观测的时候，可以用部分可观测马尔可夫决策过程来表示(Khalvati et al., 2019)。研究者在此基础上进一步提出了交互式部分可观测马尔可夫决策过程(Interactive POMDP, IPOMDP)，在该过程中每个个体的决策过程都是标准的 POMDP，即 IPOMDP 相当于 POMDP 的合集。重复信任博弈可以看成是两个个体的 IPOMDP，双方所处的状态都基于对方所做的决策以及自身形成的关于对方意图的模型(Hula et al., 2015)。

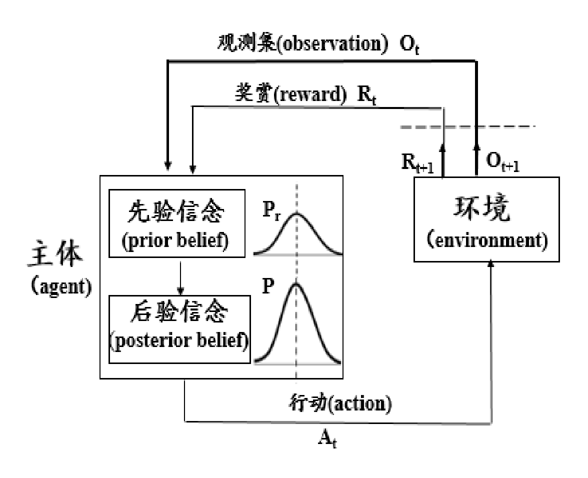


图 3. 贝叶斯模型框架图
资料来源：Friston et al.,2013

贝叶斯模型主要用于在不确定情境下人们如何基于意图推断进行决策的研究中。有研究者从信念推断入手，将“心理理论(Theory of Mind)”即人们推断自身或者他人意图的能力引入到贝叶斯模型中(Ray et al.,2009)。研究者认为人们在博弈任务中需要心理理论对玩家的意图以及行为进行策略推理(strategic reasoning)(Gonzalez & Chang, 2019; Ong et al., 2019)。在信任博弈中，心理理论体现在博弈双方对于对方是何种类型的对手进行推断的过程，比如投资者会推断被信任者是何种类型，被信任者会推断投资者是何种类型，投资者推断自身在被信任者中是何种类型等等(Rusch & Gläscher, 2019)。由此研究者根据不同被试在推断时涉及的层次将人们分成不同思维深度的被试，并用参数衡量个体的思维深度(Ray et al., 2009; Xiang et al., 2012)。Friston 等(2013)则从环境/认知的不确定性入手，将最小自由能原则(free-energy principle)引入到贝叶斯模型中，提出主观推断模型(active inference)，以

此来模拟人们在信任博弈中的决策行为(Moutoussis et al., 2014)。和基于心理理论的贝叶斯模型不同，主观推断贝叶斯模型中的参数衡量的是个体对自身策略的精确度，而不是策略涉及的思维深度。

3.2 基于强化学习模型的信任博弈行为学和影像学研究

强化学习模型能够帮助研究者更好的理解，在重复信任博弈中，投资者如何基于环境信息做决策以及决策背后的神经机制。环境信息包括被试在决策前获得的对方先验可信度信息和在互动过程中获得的对方的可信度信息(Fareri et al., 2012, 2015; Fouragnan, 2013)。

3.2.1 行为学研究

目前行为学的研究主要用强化学习模型探究健康被试在有先验可信度下，人们是如何做出信任决策的。Chang 等(2010)通过提供面孔可信度（高/中/低）的先验名誉线索，通过对比三种基于模型的强化学习模型研究了先验可信度如何影响信任的建立。这三种强化学习模型是基于损失规避(Gain and Loss)理论的模型（与获得收益相比，人们更愿意规避风险）、基于确认偏差(Confirmation Bias)理论的模型(人们在互动中，与建议与反馈结果不一致的信息相比，人们对建议与反馈结果一致的信息权重更大)，以及作者提出的动态信念迭代模型。动态信念迭代模型认为先验信息在信任博弈的整个过程中都对被试的信任行为有影响，被试会在先验信息的基础上形成对方多大可能性互惠的信念，然后该信念会随着实际的经历迭代更新。结果发现，动态信念迭代模型是预测先验可信度影响信任形成过程的最佳模型。因此，研究者认为信任是基于先验可信度的信念动态迭代而建立的。

在另一项研究中，被试首先通过掷球游戏(ball-tossing game)学习到被信任者的性格好坏（好/中/坏），之后作为投资者完成重复信任博弈 (Fareri et al., 2012)。作者使用了考虑获益损失理论的基于模型的强化学习模型。结果发现，人们习得的初始社会印象会与后续直接互动得到的反馈信号相互作用，社会印象会影响人们在互动过程中的信任行为，而互动中产生的反馈结果也会反过来会影响初始社会印象，初始印象会在重复博弈的过程中迭代更新。Fareri 等(2015)在此基础上探究了与被试亲密关系程度不同的对手如何影响其重复信任博弈中的信任行为。研究者选取了朋友、陌生人和计算机对手，探究先验可信度水平高低对信任行为的影响。结果发现人们对于先验名誉越高的对手，在互动过程中的预测偏差越小。除此之外，Radell 等(2016) 也采用了相同的实验设计以及 RW 强化学习模型探讨了不同抑制型人格（在社交中是否倾向于做出回避行为）对不同可信度水平的对手如何做信任决策。结果发现，相比于非抑制型被试，抑制型被试会对中等可信度水平对手做出的信任行为更少，这是因为抑制型被试对于中等可信度水平对手的初始信任值更低。这表明在社交中倾向于采取回避行为的被试对于中性信息或者模糊信息的解释更消极。

除了利用强化学习模型来研究先验名誉线索如何影响信任行为外，研究者也对比了有无先验信息时人们是如何做出信任决策行为的。Fouragnan (2013)提出投资者获得对手先验

可信度的方式有两种，除了提前告知对手先验可信度的有先验方式之外，还包括投资者与对手直接互动的无先验信息的方式。研究者通过对比无模型和基于模型的强化学习模型，探究了在有无先验信息条件下，被试在面对可信度水平高/低的对手做出信任行为背后的心理机制。该研究发现信念适应模型是解释投资者在重复信任博弈中信任行为的最佳模型。信念适应模型认为在重复信任博弈中，先验可信度信息作为一种社会信号，不仅会影响人们的初始决策值，还会基于互惠反馈的经验影响人们后续迭代的决策函数。人们基于先验可信度信息形成一个关于对方是否值得相信的对方可信度水平的信念(Trustworthiness belief, TW)。这种关于对方可信度水平的信念同金钱反馈结果一样作为奖励(bonus)在效用函数中进行迭代。研究发现，人们在有无先验可信度信息的两种方式下做信任决策时，首先在金钱反馈结果的基础上形成关于对手是否可信的信念，然后根据这一信念调整自身的决策，进而做出适应性的投资行为。和无先验可信度信息不同的是，有先验的条件会改变信任者对于对手是否可信的初始预期值(Fouragnan, 2013)。也有研究使用强化学习模型发现在重复信任博弈中，人们做出的信任行为是博弈双方相互学习的过程，在这个过程中，人们基于多次互动的反馈结果进行决策，而且对于多次互动中消极的结果反应更敏感，即面对消极反馈结果时，人们在下一次博弈中会快速调整自身的决策，从而做出适应性的行为(Haiyan, 2018)。

总之，行为学研究发现在重复信任博弈中，信任是一个不断学习的过程，是人们通过评估多次互动中得到的结果习得对方名誉水平然后决定是否相信对方的过程，使用强化学习理论可以揭示信任的动态建立过程。

3.2.2 影像学研究

采用功能磁共振成像技术，研究者进一步探究了先验可信度促进信任形成的神经机制。Fareri 等(2012)探究了被试在决策前获得对方的先验可信度对其信任行为以及与奖赏相关的脑功能活动的影响。在该研究中，被试先与电脑模拟的三种不同信任水平的对手（高/中/低）玩掷球游戏，以习得对手是否可信的初始印象。在接下来的重复信任博弈中，被试扮演投资者的角色分别与这些对手进行博弈。实际上对手的行为是随机的，与掷球游戏中的行为无关。研究者采用 RW 强化学习模型分析了重复信任博弈中被试的行为及其与脑活动的关联，发现与不一致情况相比，当被试所经历的对手与先验印象一致时被试更新信念的速率更快；在面对积极/消极反馈结果时，与中性反馈结果相比被试的纹状体和前扣带回的激活程度增加；而且模型中的学习速率参数与这些脑区的 BOLD 信号变化显著相关。这表明，奖赏回路脑区的 BOLD 信号反映了在行为水平上用来更新行为的预期偏差信号，说明这些脑区负责获益/受损背景下被试通过预期偏差更新信念的过程。这些结果表明，人们从直接社会互动中习得的初始印象会通过强化学习机制在一致信息的基础上，得到不断的更新。Fouragnan (2013)通过对比被试在有无先验可信度两种条件下的信任决策及相应激活的脑区，探究了先验可信度影响人们信任决策的神经基础。结果发现，纹状体激活与强化学

习模型对行为的估计只有在无先验可信度条件下才显著相关，先验可信度会打破这种相关性；在有先验可信度条件下，被试面对对手信任违规的行为时出现的纹状体负激活与被试在强化学习模型中的学习速率相关，但是在无先验可信度条件下却不相关，并且被试会持续依赖先验信息即使经验与先验不相符时。而且，和无先验条件相比，在有先验条件下合作型对手做出信任违规行为时，被试的尾状核负激活更强。先验信息会增强纹状体和腹外侧前额皮层(ventrolateral prefrontal cortex)之间的联系，进而调节被试对违规行为的容忍度。这种容忍度与被试的报复率呈负相关。同时先验可信度也会影响被试的初始信任决策，而这反映在前额叶皮层的激活上。除此之外， Fareri 等(2015)通过在强化学习模型中加入了社会价值奖赏信号探究了被试与对手的亲密程度影响其信任行为的神经基础。被试在信任博弈中扮演投资者的角色，分别与朋友、陌生人和计算机进行信任博弈。该模型认为被试从互动中获得的反馈结果除了金钱还有社会价值奖赏信号，这种社会价值奖赏信号用被试对于对手是否可信的初始感知评分来表示。研究者将这种社会奖赏信号加入 RL 的值函数中。结果表明，被试在亲密度的基础上从反馈结果中获得社会价值奖赏信号，这种社会价值奖赏信号与腹侧纹状体、内侧前额叶皮层(medial prefrontal cortex)激活程度显著相关。这表明在重复社会互动中，人们在社会价值奖赏信号的基础上进行信任决策。

这些影像学研究不仅验证了行为学计算模型研究的结果，补充发现了先验可信度会影响被试的初始信任决策，还发现了信任动态迭代的脑基础。其中，奖赏回路中的纹状体和前扣带回反映了人们通过预期偏差更新信念的过程；先验可信度对被试初始信任决策的影响反映在前额叶皮层的激活上；纹状体与前额叶皮层之间的动态联系反映了被试在博弈过程中对自身信任行为的调节。

3.3 基于贝叶斯模型的信任博弈行为学和影像学研究

贝叶斯模型在信任博弈中的应用集中于理解在重复信任博弈中投资者如何基于意图推测做出信任决策及其背后的神经机制。

3.3.1 行为学研究

Ray 等(2009)将心理理论引入贝叶斯模型，在基于贝叶斯理论的 IPOMDP 框架下对重复信任博弈中信任行为建立了一个信念层次模型(belief hierarchy model)。该模型认为，玩家知道自身是合作/不合作的类型，但并不知晓对手玩家的类型，所以是不完全信息的动态博弈。玩家关于对手是否可信的先验信念会在观察到的对方行为的基础上以贝叶斯方式进行迭代更新。玩家本身的行为也会影响其对于对手是否可信的信念。在该过程中会涉及一系列有限的信念层次：投资者认为被信任者是什么类型的人；被信任者认为投资者眼中的自己是什么样的人等等。如果玩家在多次互动结果的基础上得出对方是否可信，博弈就会达到一个主观贝叶斯纳什均衡(Bayes-Nash Equilibrium, BNE)。该模型的创新点是在 IPOMDP 框架下的贝叶斯模型中引入了策略思维水平，以此用来解释社会效用、策略水平以及先验信念对人们信任行为的影响。通过模型反转(model inversion)，可以根据被试在实

验中动态的信任行为将被试分成策略思维水平(strategic thinking)高低的不同类型的被信任者，策略思维水平高的被试投资的次数会更高。这一模型为研究人际信任中的个体差异提供了新的思路。

Hula 等(2015)使用部分可观测蒙特卡罗规划(partially observable Monte Carlo planning, POMCP)算法探究了在 IPOMDP 框架下的重复信任博弈。结果发现，投资者在博弈 10 次左右就会形成关于对手是否可信的信念，进而会做出稳定的投资行为。所以，根据投资者在前 10 次博弈中的行为便可以推断出其内部主观模型中的最优参数值，这些博弈中的行为可以确保是投资者在其自身内部模型下做出的。使用该算法还可以通过模型反转推断出被试揣测他人意图的能力。

Friston 等(2013)在基于贝叶斯理论的 IPOMDP 框架下对人们在社会互动中的决策行为建立了一个主观推断(active inference)贝叶斯模型。该模型引入了人们对先验信念的准确度参数，对决策行为进行建模，提出使用最小自由能的原则(free-energy principle)更新后验信念。 Moutoussis 等(2014)把这种模型用在信任博弈中，将效用函数(utility functions)、先验信念和结果结合起来，建立了随博弈次数的增加，投资者对他人建立信任的演化过程，并得出投资者在与对方博弈 10 次左右时，就会形成对方是否值得可信的信念。Schwartenbeck 等(2015)通过实验发现，与最大化效用的决策理论相比，主观推断下的决策理论能更好地预测人们的经济决策行为。

研究者也尝试将贝叶斯模型应用到实际问题中。Jung 等(2017)构建了医疗情境中的信任博弈来研究安慰剂镇痛效应。在该研究中，研究者对安慰剂镇痛效应建立了贝叶斯框架即将疼痛强度和疼痛评分建立似然关系，疼痛评分对应后验分布，安慰剂镇痛效应是上行感觉信号与下行疼痛预测之间的差距。由此，人们对疼痛等级的主观评定可以基于贝叶斯模型中后验分布推断所得。通过比较贝叶斯模型与线性回归模型，研究者发现先验期望会影响人们对疼痛的感知，而且贝叶斯模型可以预测人们在医疗信任博弈中的疼痛等级评定行为。

总之，基于贝叶斯推断的行为学研究得出人们在重复信任博弈中，在与对方博弈十次左右时就可以形成对方是否可信的信念，然后在此基础上做出决策。不同的人推断他人意图的能力不同，即在博弈过程中的思维水平深度不同。

3.3.2 影像学研究

Xiang 等(2012)采用功能磁共振成像技术探究推断他人意图的能力能否作为人们在信任博弈中信任行为发生偏差的客观生物标记物(objective biomarkers)。研究者采用基于心理理论的贝叶斯模型，用模型参数表征被试思维加工的深度，从而根据思维深度将被试分成高/中/低三组。结果发现，低思维深度被试的纹状体激活程度要强于高思维深度以及中等思维深度的被试，而在高思维深度被试中与心理理论相关的颞顶结合区(Temporoparietal Junction, TPJ)激活程度要强于中等以及低思维深度组被试。这表明，低思维深度的被试对

反馈结果更敏感，并主要根据反馈结果调整自身的行为，而高思维深度被试主要通过推断他人意图进行决策。Nihonsugi 等(2015)使用 fMRI 和经颅直流电刺激(transcranial directcurrent stimulation，tDCS)技术并结合计算模型，探究了推断他人意图和反馈结果对人们信任决策行为的影响在神经机制层面是否是两个分离的系统。研究者将愧疚厌恶(guilt aversion)、不公平厌恶(inequity aversion)与强化学习模型中的效用函数结合起来建立了模型。将该模型中的愧疚敏感性参数和不公平敏感性参数与影像结果联系起来发现，右背外侧前额叶皮质(right dorsolateral prefrontal cortex, DLPFC)的激活与基于意图的经济决策行为有关，腹侧纹状体和杏仁核的激活与基于反馈的经济决策行为有关。而且对 DLPFC 的选择性刺激会增强基于意图的决策行为。这些结果表明，右侧 DLPFC 在加工实施基于意图的合作行为中起重要作用。综合其研究发现，Nihonsugi 等(2015)提出在重复信任博弈中主要包括推断他人意图和基于反馈结果做决策的两个分离的神经系统。人们通过多次互动得到的奖赏信号（纹状体的激活），推断习得对方的可信度水平（DLPFC 和扣带回的激活），然后在此基础上做出适应性行为（这些脑区的 BOLD 信号与模型中的预测偏差显著相关）；而且先验可信度会加强这两个系统之间的联系。

这些影像学研究，不仅发现了不同思维深度个体进行信任决策时个体差异的神经基础，而且发现在重复信任博弈中，存在推断他人意图和基于反馈结果做决策的两个分离的神经系统。

4 不足与展望

综上所述，采用基于计算模型的行为学和脑影像学研究方法，研究者从心理和神经机制层面发现了先验可信度和对方意图的推测是如何促进信任形成的，对“信任是如何形成的”这一问题获得了更深入的理解。但还存在一些不足和值得进一步探索的方向。

4.1 计算模型的发展

目前应用在信任博弈中的计算模型主要包括强化学习模型和贝叶斯模型。强化学习模型是基于个体完全理性的假设，认为个体在当前预期偏差的基础上更新对行为值的期望，并用学习率衡量个体对反馈结果的权重大小（学习率具有个体差异性）。该模型自提出后被广泛应用在各种学习任务中，并且研究者将其与脑影像技术结合在探究大脑的奖赏功能方面获得了许多重要发现(Jaafra et al., 2019; Lee et al., 2012)。但是强化学习模型的客观自适应的学习过程很难应用在实际生活中，因为在实际生活中个体所面对的情景是不确定的，同时行为的效用值是未知的，需要个体对其进行推断(Mathys et al., 2011)。

而贝叶斯模型基于个体有限理性的假设，结合贝叶斯理论，利用条件概率将个体的信念与所作的行为建立起联系，可以很好地刻画个体在面对不确定情景下的决策行为(Mathys et al., 2011)。其中，主观推测模型已经被尝试应用于不同领域的研究中(Friston et al., 2016; Parr & Friston, 2017; Smith et al., 2019)。研究者通过数据模拟发现了影响不同行为的特定参

数，如策略深度、决策不确定性、先验信念等等(Smith et al., 2019)。但目前仅有一项研究将其用到信任博弈中(Moutoussis et al., 2014)。未来需要更多研究应用该模型来模拟信任形成过程，确定影响信任形成的关键参数，并设计功能影像任务来检验关键参数的神经基础从而确定信任形成个体差异的心理机制和神经基础。

近年来研究者也在试图发展其它类型的计算模型。比如，Mathys 等提出了分层高斯过滤模型(Hierarchical Gaussian Filter, HGF)。该模型将贝叶斯基于概率论刻画不确定性的方式与强化学习模型中刻画个体差异的更新方式结合起来。该模型的更新方程与强化学习模型的类似（以预期误差驱动），不同之处在于分层高斯过滤模型是以个体对策略的精确度的权衡作为学习率，以此来刻画人们如何在环境以及感知的不确定下进行决策的过程，而且有研究已将该模型成功用在了需要推断他人意图的社会交换的情景中(Diaconescu et al., 2014)。除此之外，有研究将强化学习模型中的效用函数与贝叶斯模型中的用概率刻画不确定性结合起来提出了 Fehr–Schmidt 不公平厌恶模型(Fehr–Schmidt inequality aversion model, FS model)。在该模型中既有刻画个体差异的学习速率参数、不公平厌恶参数，也有个体推断他人意图涉及的思维深度参数、提前计划步骤多少的参数，可以全面的模拟人们在信任博弈中的决策行为。研究者通过比较不同干预被试组（是否接受高质量早期教育）在模型中的参数，结果发现相比于未接受过高质量早期教育的被试，接受过的被试在信任博弈等社会互动中决策时会计划更多的步骤。这表明高质量的早期教育会对人们的社会决策行为有长远有益的影响(Luo et al., 2018)。

未来可以根据研究目的，灵活地选择和发展各类模型用于信任博弈研究中，以期加深对信任动态迭代过程的理解，并促进对信任博弈中个体差异的心理和神经机制的理解。

4.2 基于计算模型的脑-行为因果关系研究

虽然基于计算模型的脑影像研究，可以从时间和空间两个维度来刻画大脑在执行某个特定认知过程时神经活动随时间的动态变化，将认知和脑功能建立起联系，解决大脑如何执行某一认知功能的问题，但是当前研究仍然不能回答行为与脑之间的因果关系。对脑损伤患者的异常决策行为的计算模型研究(Gu et al., 2015)，在推测特定脑区在认知过程中的独特作用具有重要意义，但这类研究不易被重复。以经颅磁刺激(Transcranial Magnetic Stimulation, TMS)和 tDCS 技术为代表的非侵入性脑刺激技术的出现，为探究脑与决策行为的因果关系提供了可能(荣悦彤 等, 2019)。例如， Zheng 等(2017)使用 tDCS 增强右侧 DLPFC 的兴奋性，发现并不会影响人们在信任博弈中做出的信任行为。目前仅有一项研究使用了基于模型的 fMRI 和 tDCS 技术，发现人们在作信任决策时推断他人意图和加工反馈结果是基于两个分离的神经系统(Nihonsugi et al., 2015)。未来需要更多的研究将非侵入性脑刺激技术、fMRI 技术和计算模型相结合，在信任博弈的框架下，进一步揭示信任形成过程背后的心理机制与其神经基础之间的因果关系。

4.3 精神疾病患者的人际信任研究

近年来，计算神经科学的发展促进了计算模型在临床研究中的应用，并由此发展出一个新的研究领域，即计算精神病学(computational psychiatry)(Huys et al., 2011; Montague et al., 2012; Stephan & Mathys, 2014)。计算精神病学采用基于模型的定量指标来推测异常行为和神经活动背后隐藏的原因，从而解释精神病理(Friston et al., 2014)。

以往研究发现精神疾病患者在信任博弈中会做出异常的信任行为。例如边缘性人格障碍患者(Borderline Personality Disorder, BPD)和自闭症(Autism Spectrum Disorder , ASD)儿童会表现出更少的信任行为(King-Casas. et al., 2008; Knoch et al., 2009; Maurer et al., 2018)，以及青少年抑郁症患者会表现出过度的信任行为而成年抑郁症患者则表现出更少的信任行为(Mellick et al.,2019; Wehebrink et al., 2018)。但是这些研究只是发现了患者信任行为异常，对于患者为何做出异常信任决策的心理过程和神经机制仍不清楚。目前仅有一项以精神疾病患者为研究对象的信任博弈计算模型研究。该研究采用心理理论的贝叶斯模型发现，边缘性人格障碍患者作为投资者进行博弈时会表现出不同于健康被试作为投资者的思维深度分布。该研究表明这种基于心理理论的贝叶斯模型得出的思维深度所对应的神经反应类型可以作为识别异常信任行为的客观标记物(Xiang et al., 2012)。未来的研究可以从计算精神病学视角下深入研究精神疾病患者人际信任过程建立的异常之处。将信任博弈、计算模型以及神经影像技术结合起来的研究方法，不仅可以加深我们正常状态下人们信任形成过程的理解(Sanfey, 2007)，也为研究精神障碍患者社会功能障碍提供了新的视角。

参考文献：

Anderhub, V., Engelmann, D., & Güth, W. (2002). An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior & Organization, 48*(2), 197-216.

Ben-Ner, A., Putterman, L., & Ren, T. (2011). Lavish returns on cheap talk: Two-way communication in trust games. *Journal of Socio-Economics, 40*(1), p.1-13.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122-142.

Blue, P. R., Hu, J., Peng, L., Yu, H., Liu, H., & Zhou, X. (2020). Whose promises are worth more? How social status affects trust in promises. *European Journal of Social Psychology*, *50*(1), 189-206.

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87-105. doi:10.1016/j.cogpsych.2010.03.001

Charpentier, C. J., & O' Doherty, J. P. (2018). The application of computational models to social neuroscience: Promises and pitfalls. *Social Neuroscience, 13*(6), 637-647.

Cheong, J. H., Jolly, E., Sul, S., & Chang, L. J. (2017). Computational models in social neuroscience. *Computational Models of Brain and Behavior*, 229-244.

Cisler, J. M., Bush, K., Scott Steele, J., Lenow, J. K., Smitherman, S., & Kilts, C. D. (2015). Brain and behavioral evidence for altered social learning mechanisms among women with assault-related posttraumatic stress disorder. *Journal of Psychiatric Research, 63*, 75-83. doi:10.1016/j.jpsychires.2015.02.014

Claus, C., & Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI, 1998*(746-752), 2.

Cochard, F., Nguyen Van, P., & Willinger, M. (2004). Trusting behavior in a repeated investment game. *Journal of Economic Behavior & Organization, 55*(1), 31-44. doi:10.1016/j.jebo.2003.07.004

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology, 16*(2), 199-204.

Declerck, C. H., Boone, C., & Emonds, G. (2013). When do people cooperate? The neuroeconomics of prosocial decision making. *Brain and Cognition, 81*(1), 95-117.

Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., . . . Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology, 10*(9).

Engelmann, J. B. (2010). Measuring Trust in Social Neuroeconomics: a Tutorial. *Hermeneutische Blätter, 1*(2), 225-242.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6*, 148. doi:10.3389/fnins.2012.00148

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *Journal of Neuroscience, 35*(21), 8170-8180. doi:10.1523/JNEUROSCI.4775-14.2015

Fehr, E., & Schmidt, K. M. (2005). The economics of fairness, reciprocity and altruism‑Experimental evidence and new theories. *Economics, 20*, 51.

Fett, A. K., Gromann, P. M., Giampietro, V., Shergill, S. S., & Krabbendam, L. (2014). Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Social Cognitive and Affective Neuroscience, 9*(4), 395-402. doi:10.1093/scan/nss144

Fouragnan, E. (2013). The neural computation of trust and reputation. *Biochemistry*, 52(29), 4941-54.

Friston, K., Fitzgerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Review, 68*, 862-879.

Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers Humman Neuroscience, 7*, 598. doi:10.3389/fnhum.2013.00598

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry, 1*(2), 148-158.

Gonzalez, B., & Chang, L. J. (2019). Computational models of mentalizing. *PsyArXiv*. doi:doi:10.31234/osf.io/4tyd9

Gu, X., Wang, X., Hula, A., Wang, S., Xu, S., Lohrenz, T. M., . . . Montague, P. R. (2015). Necessary, yet dissociable contributions of the insular and ventromedial prefrontal cortices to norm adaptation: Computational and lesion evidence in humans. *Journal of Neuroscience, 35*(2), 467-473. doi:10.1523/JNEUROSCI.2906-14.2015

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology, 24*, 92-97.

Haiyan, L. (2018). Dynamic trust game model between venture capitalists and entrepreneurs based on reinforcement learning theory. *Cluster Computing*. doi:10.1007/s10586-017-1666-x

Hula, A., Montague, P. R., & Dayan, P. (2015). Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS Computational Biology, 11*(6).

Huys, Q. J., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? , *Neural Networks, 24*(6), 544-551.

Jaafra, Y., Laurent, J. L., Deruyver, A., & Naceur, M. S. (2019). Reinforcement learning for neural architecture search: A review. *Image and Vision Computing, 89*, 57-66.

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology, 32*(5), 865-889.

Jung, W. M., Lee, Y. S., Wallraven, C., & Chae, Y. (2017). Bayesian prediction of placebo analgesia in an instrumental learning model. *PLoS One, 12*(2), e0172609. doi:10.1371/journal.pone.0172609

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1995). *Partially observable Markov decision processes for artificial intelligence.* Paper presented at the International Workshop on Reasoning with Uncertainty in Robotics.

Keser, C. (2003). Experimental games for the design of reputation management systems. *IBM Systems Journal, 42*(3), 498-506.

Khalvati, K., Park, S. A., Mirbagheri, S., Philippe, R., & Rao, R. (2019). Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science Advances, 5*(11), eaax8783.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science, 308*(5718), 78-83.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science, 321*(5890), 806-810. doi:10.1126/science.1156902

Knoch, D., Schneider, F., Schunk, D., Hohmann, M., & Fehr, E. (2009). Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. *Proceedings of the National Academy of Sciences, 106*(49),

20895-20899.

Kreps, D. M. (1990). Corporate culture and economic theory. *Perspectives on Positive Political Economy, 90*, 109-110.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., . . . Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences, 104*(50), 20084-20089.

Krueger, F., & Meyer-Lindenberg, A. (2019). Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends in Neurosciences, 42*(2), 92-101.

Lee, D., Seo, H., & Jung, M.(2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience, 35*, 287-308.

Luo, Y., Hétu, S., Lohrenz, T., Hula, A., & Ramey, C. (2018). Early childhood investment impacts social decision-making four decades later. *Nature Communications, 9*(1).

Ma, Q., Liang, M., Qiang, S., & Yu, R. (2015). You have my word: Reciprocity expectation modulates feedback-related negativity in the Trust Game. *Plos One, 10*(2), e0119129-.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience, 5*, 39.

Maurer, C., Chambon, V., Bourgeois-Gironde, S., Leboyer, M., & Zalla, T. (2018). The influence of prior reputation and reciprocity on dynamic trust-building in adults with and without autism spectrum disorder. *Cognition, 172*, 1-10. doi:10.1016/j.cognition.2017.11.007

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization, 52*(2), 267-275. doi:10.1016/s0167-2681(03)00003-9

Mellick, W., Sharp, C., & Ernst, M. (2019). Depressive adolescent girls exhibt atypical social decision-making in an iterative trust game. *Journal of Social and Clinical Psychology, 38*(3), 224-244. doi:10.1521/jscp.2019.38.2.224

Montague, P. R. (2018). Computational Phenotypes Revealed by Interactive Economic Games. In A. Anticevic & J. D. Murray (Eds), Computational psychiatry: Mathematical modeling of mental illness (pp. 273-292): Academic Press

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences, 16*(1), 72-80.

Moutoussis, M., Trujillo-Barreto, N. J., El-Deredy, W., Dolan, R., & Friston, K. (2014). A formal model of interpersonal inference. *Frontiers in Human Neuroscience, 8*, 160.

Nihonsugi, T., Ihara, A., & Haruno, M. (2015). Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. *Journal of Neuroscience, 35*(8), 3412-3419. doi:10.1523/JNEUROSCI.3885-14.2015

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model‐based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences, 1104*(1), 35-53.

Ong, D. C., Zaki, J., & Goodman, N. D. (2019). Computational Models of Emotion Inference in Theory of Mind: A Review and Roadmap. *Topics in Cognitive Science, 11*(2), 338-357.

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific Reports, 2017, 7*(1):1-21.

Premack, D., & Woodruff, G. (1978). Does a chimpanzee have a theory of mind. *Behavioral & Brain Sciences, 1*(4), 515-526.

Puterman, M. L. (1995). Markov decision processes: Discrete stochastic dynamic programming. *Journal of the Operational Research Society, 46*(6), 792-792.

Radell, M. L., Sanchez, R., Weinflash, N., & Myers, C. E. (2016). The personality trait of behavioral inhibition modulates perceptions of moral character and performance during the trust game: Behavioral results and computational modeling. *PeerJ, 4*, e1631. doi:10.7717/peerj.1631

Ray, D., King-Casas, B., Montague, P. R., & Dayan, P. (2009). Bayesian model of behaviour in economic games. In D. Koller, D. Schuurmans, Y. Bengio & L. Bottou. (Eds), Advances in neural information processing systems 21 (pp. 1345-1352).

Riedl, R., & Javor, A. (2012). The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging. *Journal of Neuroscience, Psychology, and Economics, 5*(2), 63-91. doi:10.1037/a0026318

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality, 35*(4), 651-665.

Rusch, T., & Gläscher, J. (2019). Classification of Theory of Mind tasks and their computational models. *PsyArXiv, 7*. doi:10.31234/osf.io/uf85z

Sanfey, A. G. (2007). Social Decision-Making: Insights from Game Theory and Neuroscience. *Science, 318*.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Kronbichler, M., & Friston, K. (2015). Evidence for surprise minimization over value maximization in choice behavior. *Scientific Reports, 5*, 16575. doi:10.1038/srep16575

Smith, R., Khalsa, S. S., & Paulus, M. P. (2019). An Active Inference Approach to Dissecting Reasons for Nonadherence to Antidepressants. *Biological Psychiatry:Cognitive Neurosci ence and Neuroimaging*. doi:10.1016/j.bpsc.2019.11.012

Smith, R., Lane, R. D., Parr, T., & Friston, K. J. (2019). Neurocomputational mechanisms underlying emotional awareness: Insights afforded by deep active inference and their potential clinical relevance. *Neuroscience & Biobehavioral Reviews, 107*, 473-491. doi:10.1016/j.neubiorev.2019.09.002

Snijders, C., & Keren, G. (2001). Do You Trust? Whom Do You Trust? When Do You Trust? , *Advances in Group Processes, 18*(18), 129-160.

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology, 25*, 85-92.

Tzieropoulos, H. (2013). The Trust Game in neuroscience: A short review. *Social Neuroscience, 8*(5), 407-416. doi:10.1080/17470919.2013.832375

Wehebrink, K. S., Koelkebeck, K., Piest, S., de Dreu, C. K. W., & Kret, M. E. (2018). Pupil mimicry and trust - Implication for depression. *Journal of Psychiatric Research, 97*, 70-76. doi:10.1016/j.jpsychires.2017.11.007

Wilson, R. K., & Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly, 59*(2), 189-202.

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Computtational Biology, 8*(12), e1002841. doi:10.1371/journal.pcbi.1002841

Zheng, H., Wang, S., Guo, W., Chen, S., Luo, J., Ye, H., & Huang, D. (2017). Enhancing the activity of the DLPFC with tDCS alters risk preference without changing interpersonal trust. *Frontiers in Neuroscience, 11*, 52.

Chen X., & Ye H. S. (2009). About the research on trust in the perspective of behavioral theory [行为博弈视野下信任研究的回顾]. *Psychological Science, 32*(3), 636-639.

Chen Y., Xu M. X., & Wang J. X, (2020). The cognitive neural network model of trust [信任的认知神经网络模型]. Advances in *Psychological Science*. *28*(05), 800-809.

Rong Y. T,, Wang X. M,, & Zhou Y, (2019). Application of non-invasive brain stimulation techniques in decision-making behavior [非侵入性脑刺激技术在决策行为研究中的应用]. *Chinese Journal of Behavioral Medicine and Brain Science*

Shi Y. W,, Xu F. M,, Luo J. J,, Li Y,, & Liu C.H, (2015). Trust in behavioral economics: Formation mechanisms and influential factors [行为经济学中的信任: 形成机制及影响因素]. *Advances in Psychological Science*. *23*(7), 1236-1244.

Zhang N,, Zhang Y. Q,, & Wu K. K. (2011). Psychological and Neurophysiologic Mechanisms of Trust [信任的心理和神经生理机制]. *Psychological Science*. *34*(5), 1137-1143.

Zhang W,, Zhang Z,, Gao Y,, Duan H. P,, & Wu X. N, (2016). The theoretical models and brain mechanisms of interpersonal trust game during economic decision-making [经济决策中人际信任博弈的理论模型与脑机制]. *Advances in Psychological Science*. *24*(11), 1780-1791.

Psychological and neural mechanisms of trust formation：A perspective from computational modeling based on the decision of investor in the trust game

GAO Qinglin[1,2]; ZHOU Yuan[1,2]
([1]*Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China*)
([2]*Department of Psychology, University of Chinese Academy of Sciences, Beijing, 100049, China)*

Abstract： Interpersonal trust has permeated all aspects of social exchange. It is the foundation of promoting and maintaining social corporation. Using the trust game paradigm, previous studies have investigated the theoretical models, biological bases and influential factors of interpersonal trust. In recent years, computational modeling has been increasingly applied to the research field of interpersonal trust. It enables researchers to explore the psychological mechanisms underlying interpersonal trust. Combining computational modeling with neuroimaging technology can deepen our understanding of the brain mechanisms of trust behaviors. The current application of the computational modeling to the trust game primarily aimed to answer the question of "how trust is formed". Future researchers could further combine advanced computational modeling techniques with non-invasive brain stimulation technologies to uncover the unique process of trust formation among patients with mental disorders. By doing so, we hope to gain a better understanding about the differences in the psychological and neural mechanisms of trust formation between healthy population and patients with mental disorders.

Key Words： interpersonal trust; trust game; computational modeling; functional magnetic resonance imaging